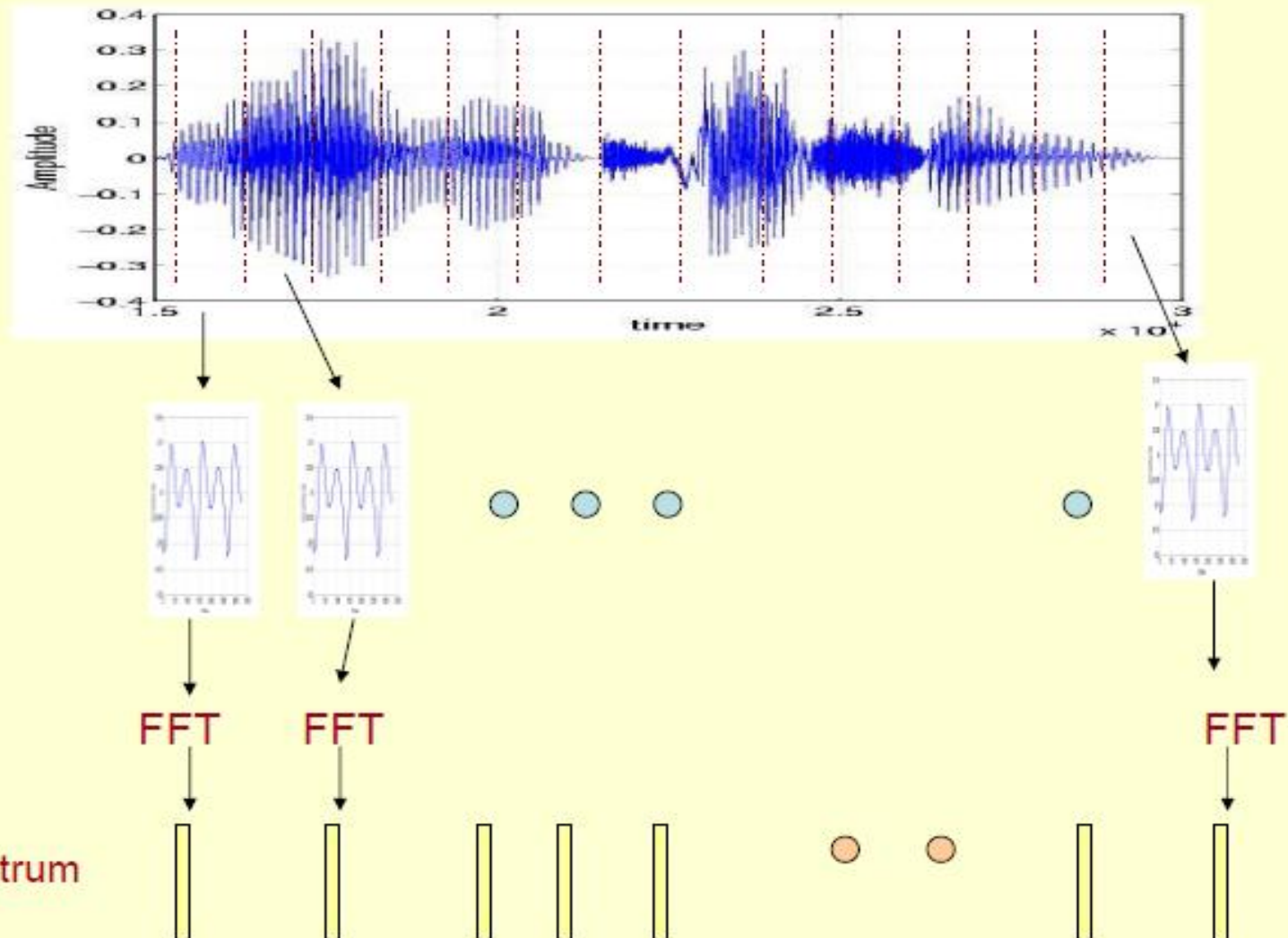
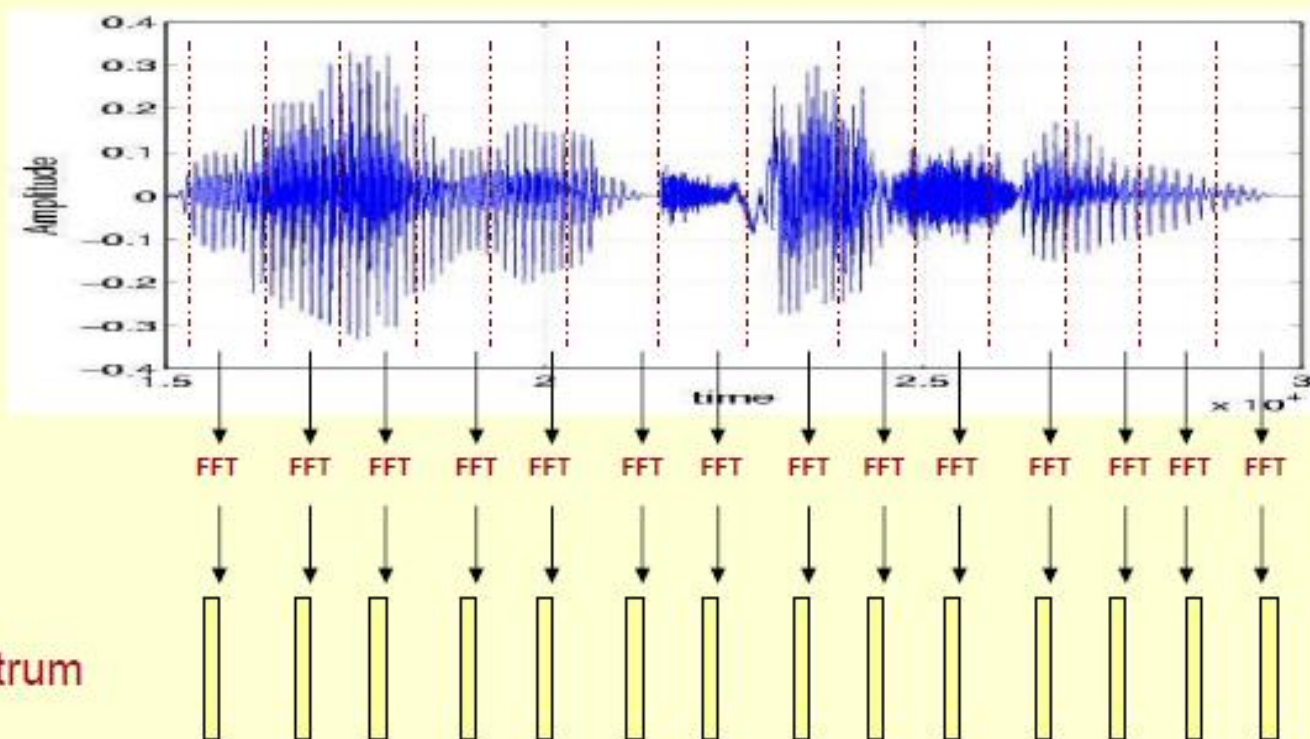


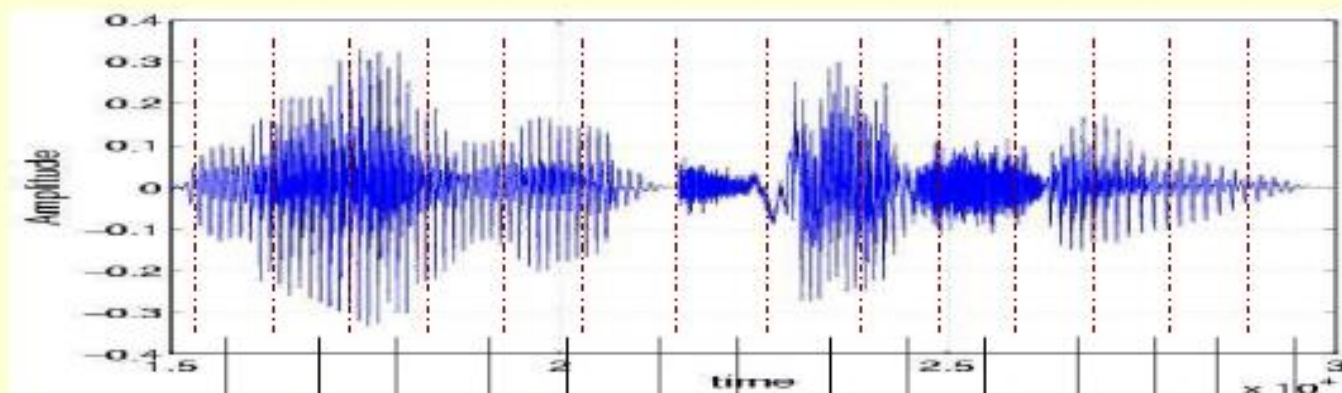
Speech signal represented as a sequence of spectral vectors



Speech signal represented as a sequence of spectral vectors

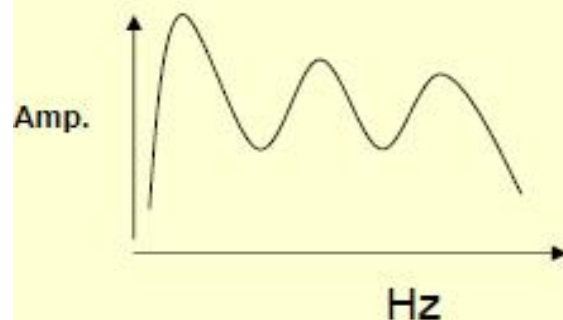
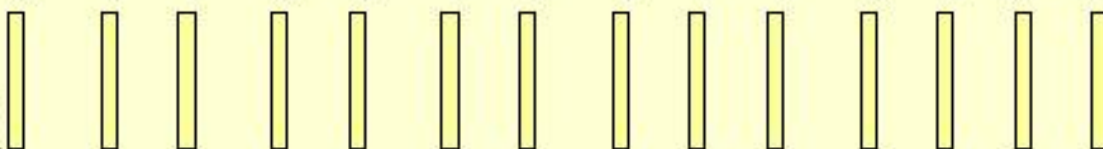


Speech signal represented as a sequence of spectral vectors

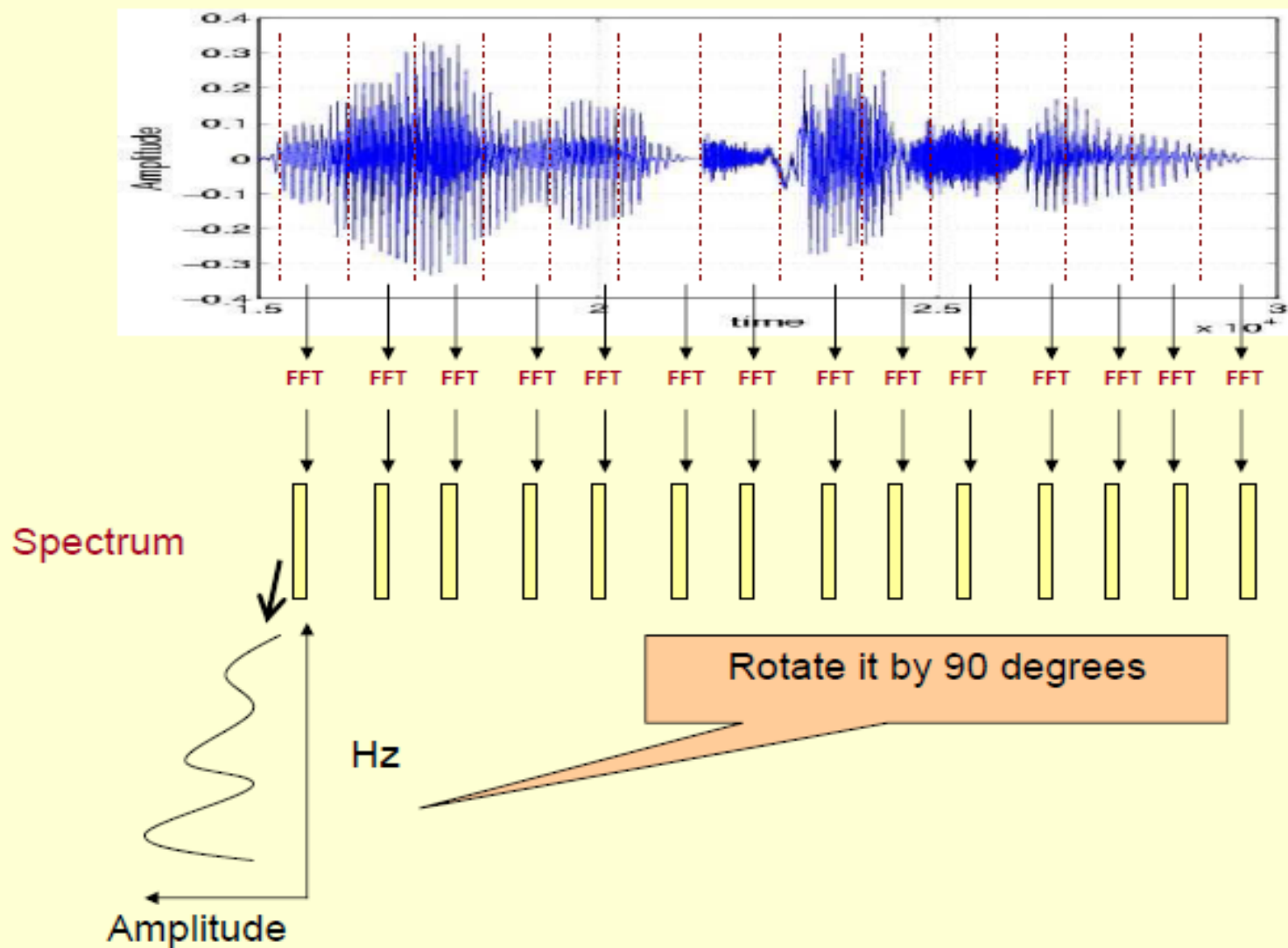


FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT

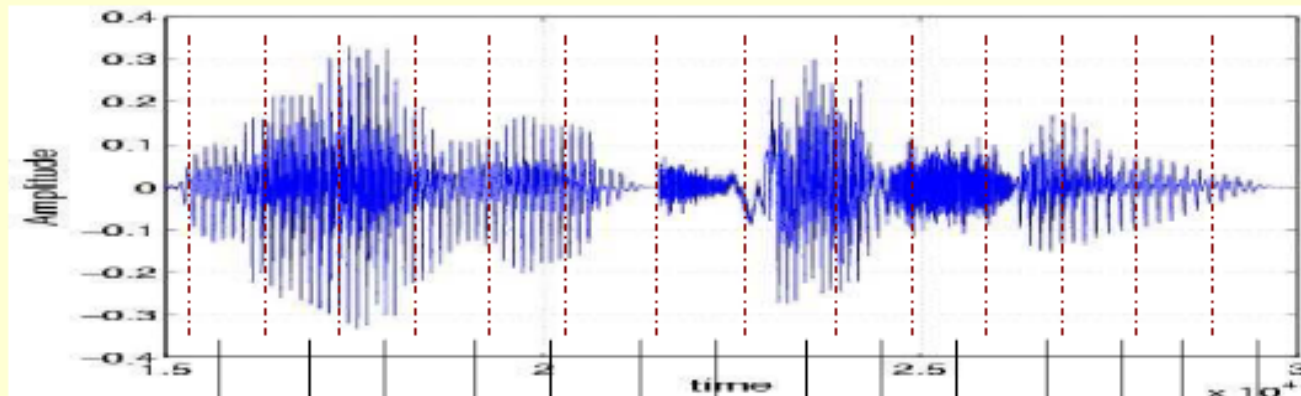
Spectrum



Speech signal represented as a sequence of spectral vectors

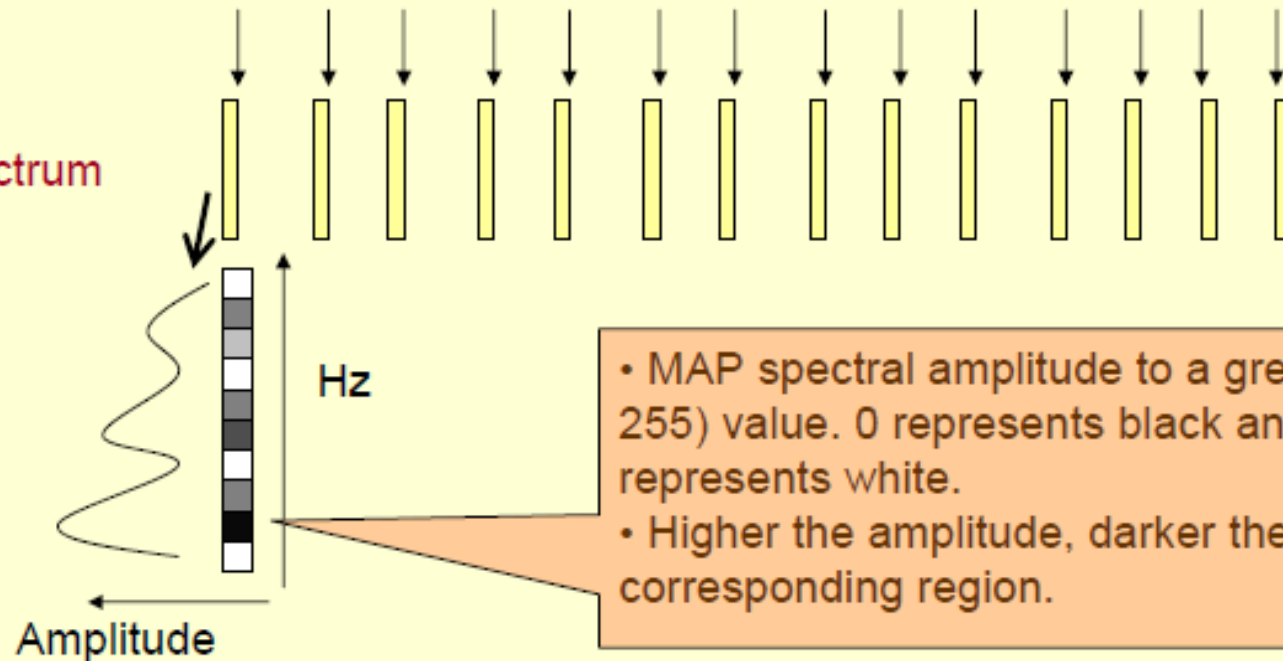


Speech signal represented as a sequence of spectral vectors



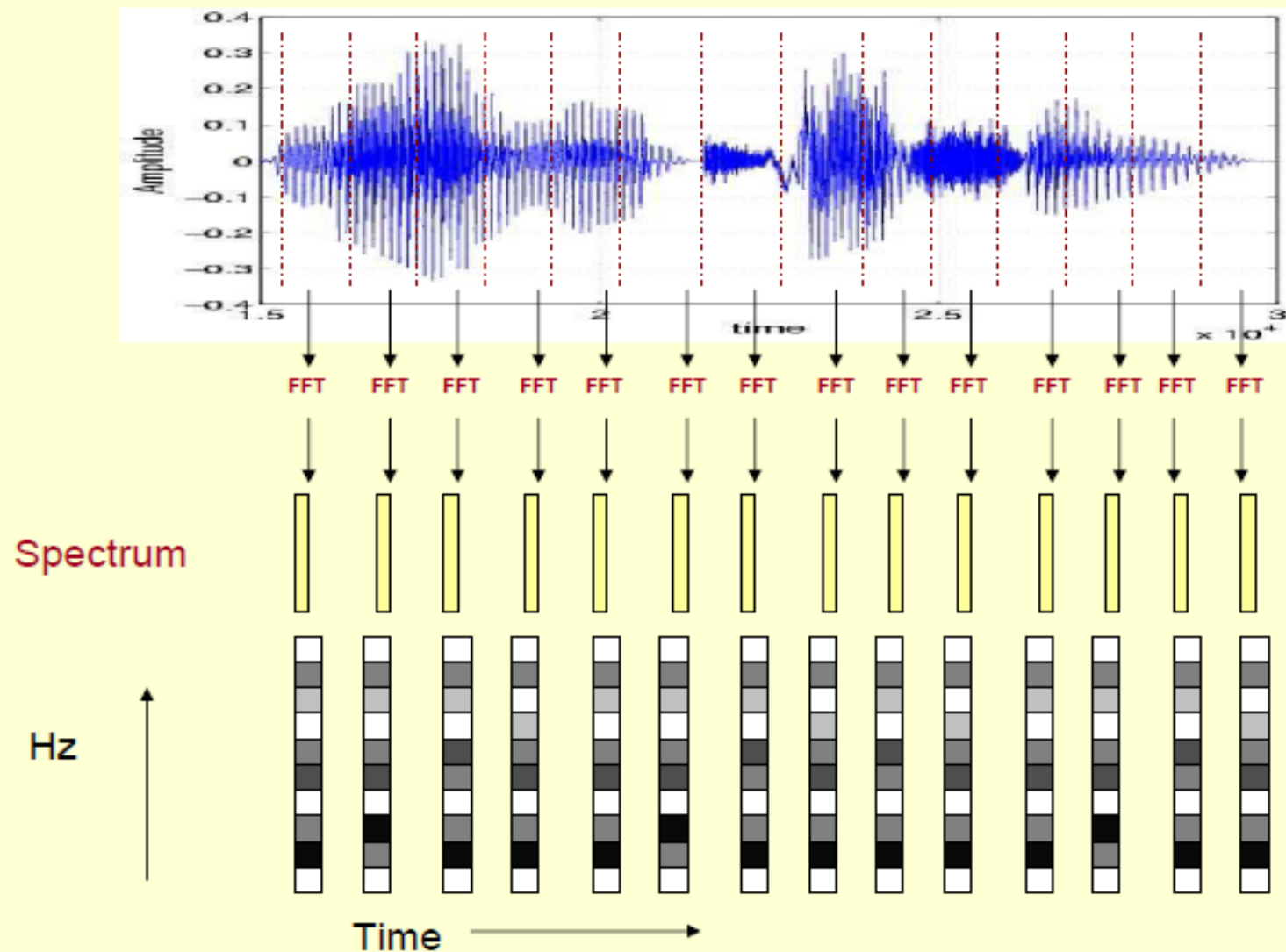
FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT

Spectrum

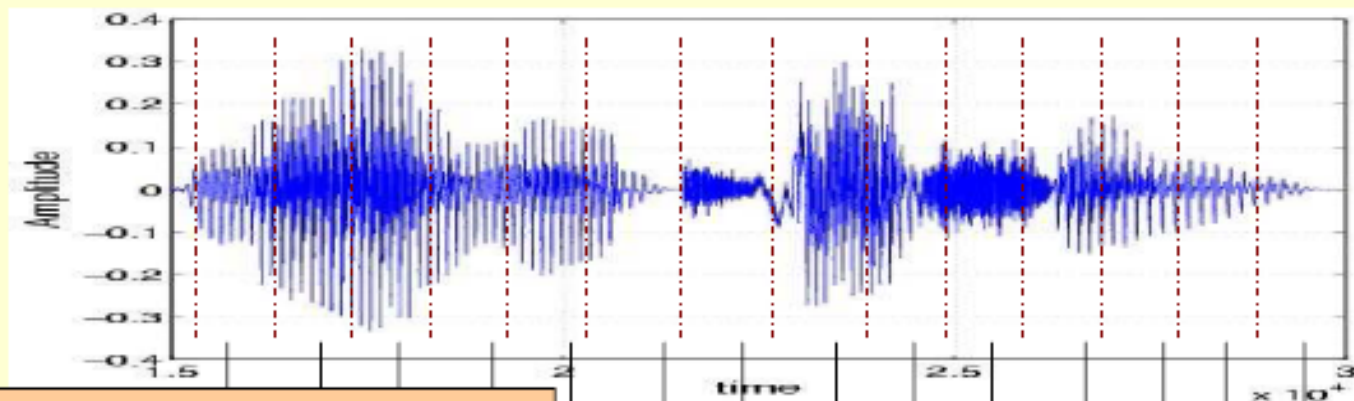


- MAP spectral amplitude to a grey level (0-255) value. 0 represents black and 255 represents white.
- Higher the amplitude, darker the corresponding region.

Speech signal represented as a sequence of spectral vectors



Speech signal represented as a sequence of spectral vectors



Time Vs Frequency representation of a speech signal is referred to as spectrogram

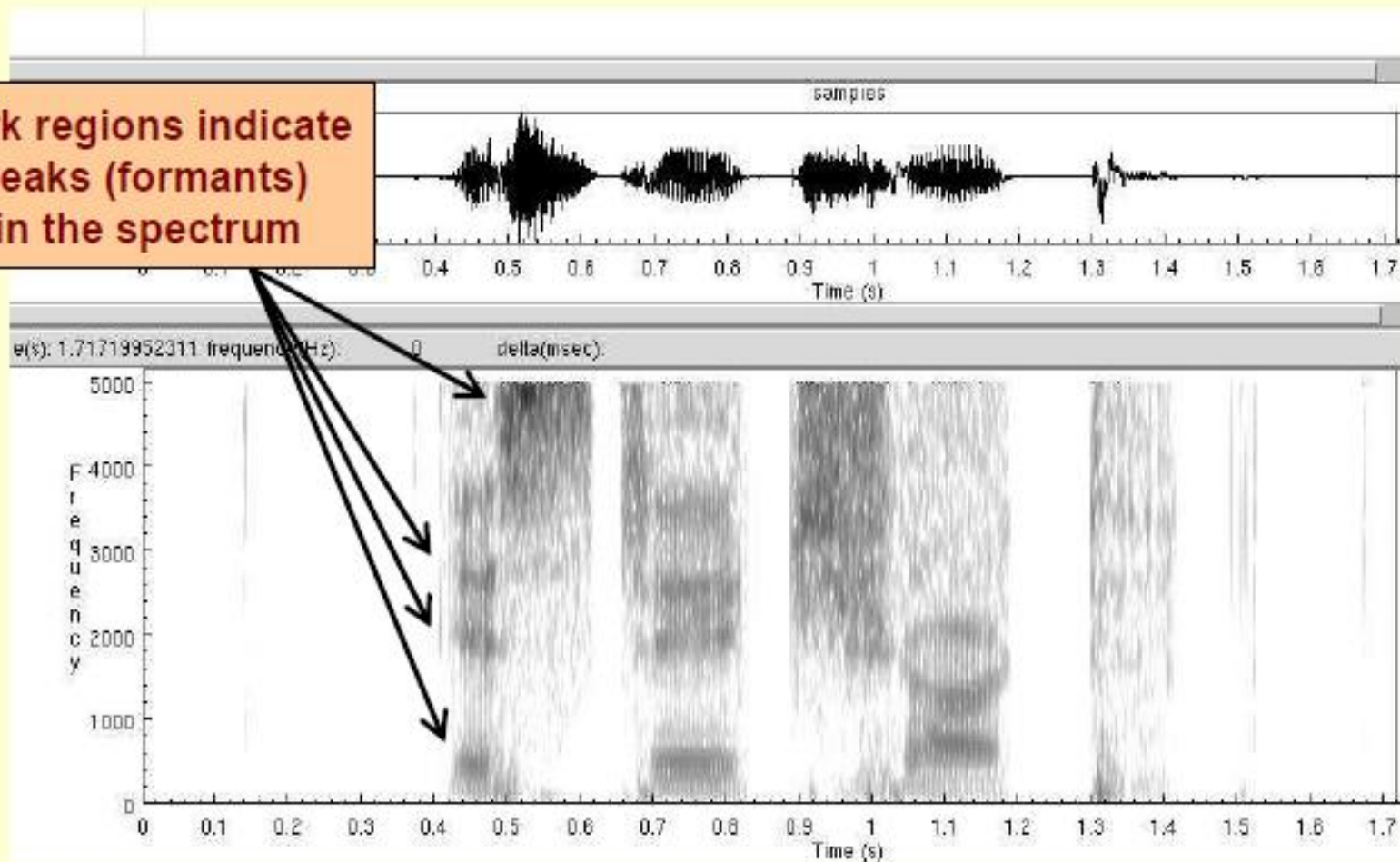
FFT FFT FFT FFT FFT FFT FFT FFT FFT FFT

Hz

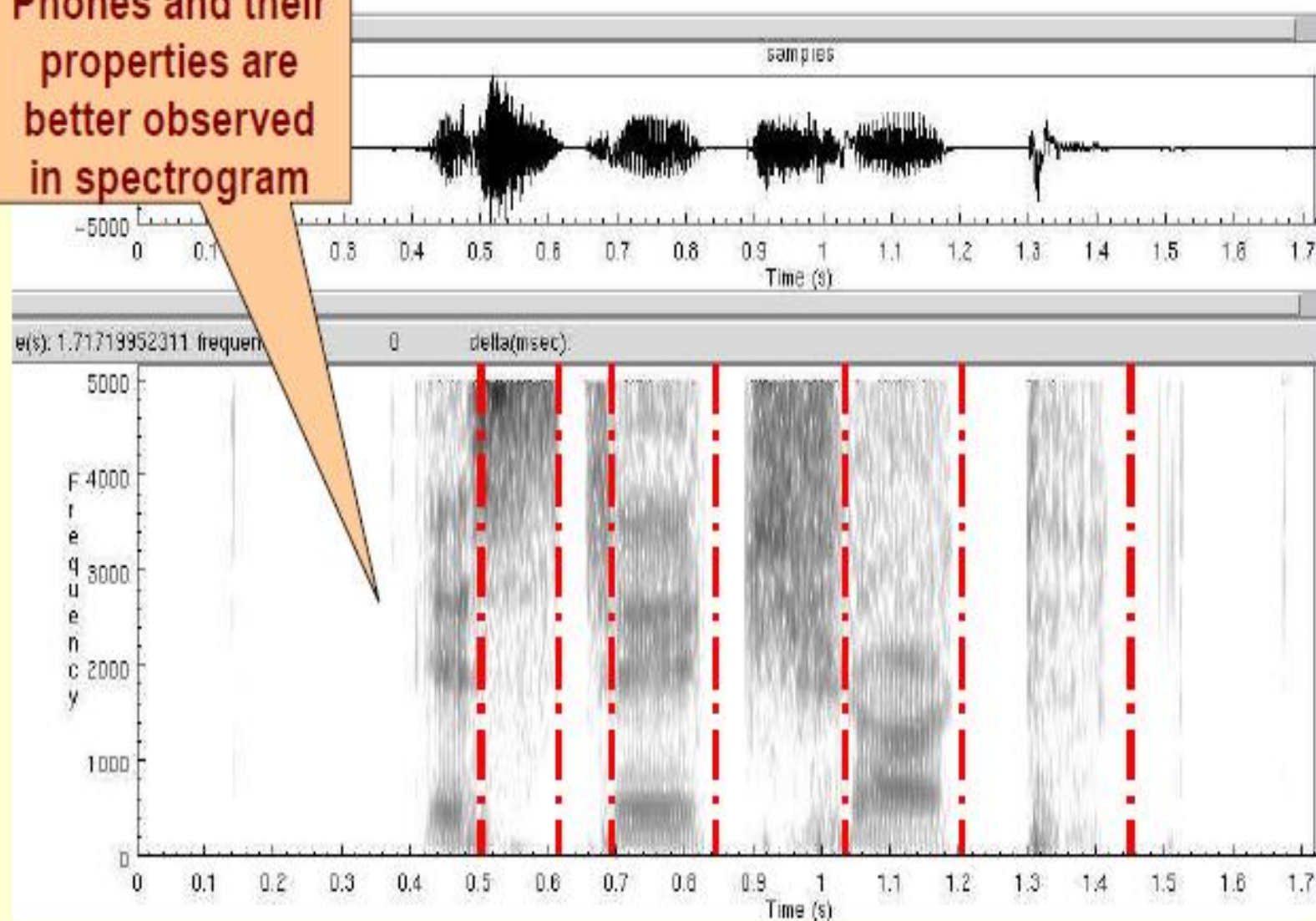
Time

Some Real Spectrograms

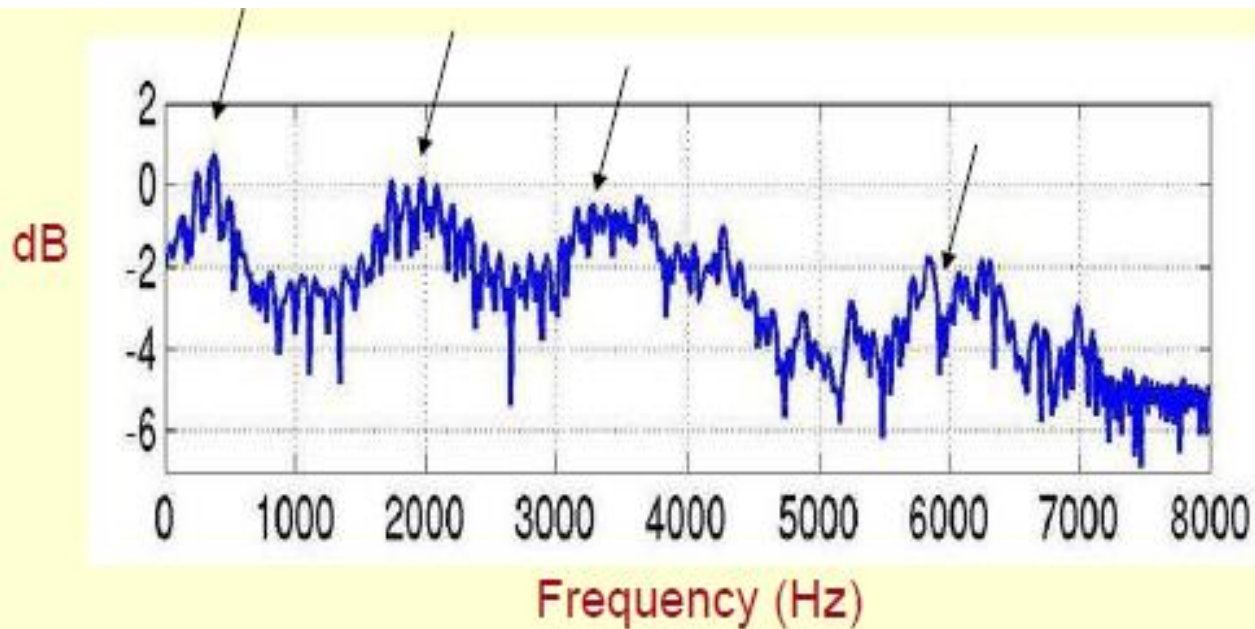
Dark regions indicate
peaks (formants)
in the spectrum



Phones and their
properties are
better observed
in spectrogram



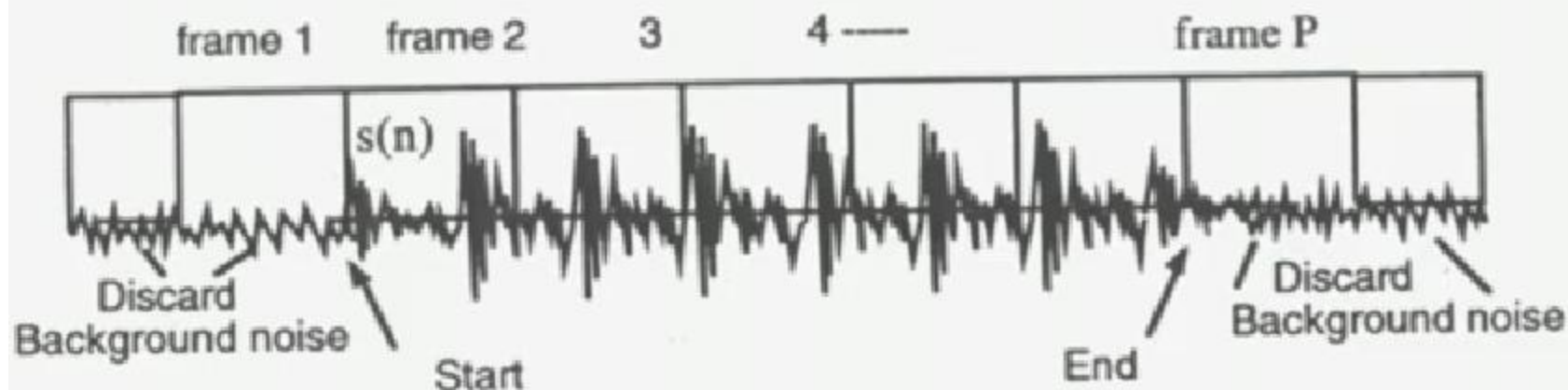
- Time-Frequency representation of the speech signal
- Spectrogram is a tool to study speech sounds (phones)
- Phones and their properties are visually studied by phoneticians
- Hidden Markov Models implicitly model spectrograms for speech to text systems
- Useful for evaluation of text to speech systems
 - A high quality text to speech system should produce synthesized speech whose spectrograms should nearly match with the natural sentences.



- Peaks denote dominant frequency components in the speech signal
- Peaks are referred to as formants
- Formants carry the identity of the sound

The major advantage of the all-pole model is that the gain parameter G and the filter coefficients (a_k , $k = 1, 2, 3, \dots$) can be easily estimated in a very straightforward and computationally efficient manner, also with good accuracy.

Any given utterance will last a certain amount of time. It is split into frames for processing as given:



Basic Parameter Extraction

- There are a number of very basic speech parameters which can be easily calculated for use, in simple applications:
 - Short Time Energy
 - Short Time Zero Cross Count (ZCC)
 - Pitch Period
- All of the above parameters are typically estimated for frames of speech between 10 and 20 ms long

Short Time Energy

- The short-time energy of speech may be computed by dividing the speech signal into frames of N samples and computing the total squared values of the signal samples in each frame.
- Splitting the signal into frames can be achieved by multiplying the signal by a suitable window function $w(n)$ $\{n=0, 1, 2, 3, \dots, N-1\}$, which is zero for n outside the range $(0, N-1)$

Rectangular Window

- A simple rectangular window of duration of 12.5 ms is suitable for this purpose. For a window starting at sample m , the short-time energy E_m is defined as

$$E_m = \sum_n [s(n) w(m-n)]^2$$

$$w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases}$$



$$E_m = \sum_n [s(n)]^2 h(m-n)$$

$$h(n) = [w(n)]^2$$

Linear filter representation

- The above equation (see previous slide) can thus be interpreted as



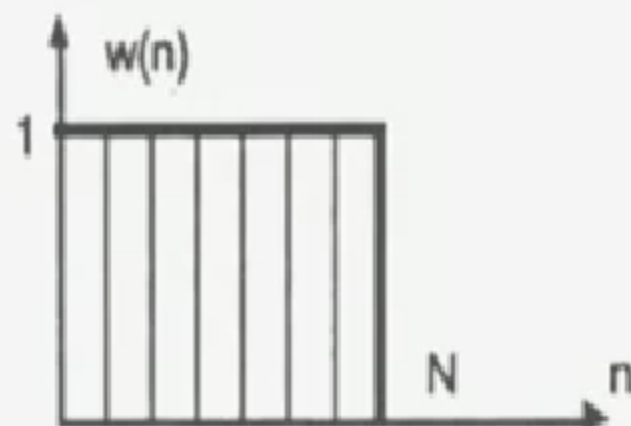
The signal $s(n)^2$ is filtered by a linear filter with impulse response $h(n)$.

The choice of the impulse response, $h(n)$ or equivalently the window, determines the nature of the short-time energy representation.

To see how the choice of window affects the short-time energy, let us observe that if $h(n)$ was very long and of constant amplitude E_m would change very little with time

Such a window would be equivalent of a very narrowband lowpass filter. Clearly what is desired is some lowpass filtering, so that the short-time energy reflects the amplitude variations of the speech signal.

Note: Rectangular window



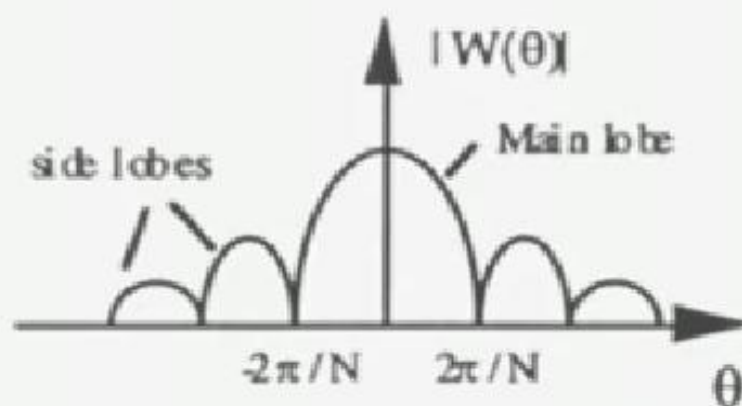
$$w(n) = \begin{cases} 1 & 0 \leq n \leq N-1 \\ 0 & \text{otherwise} \end{cases}$$

$$W(z) = \sum_{n=0}^{N-1} w(n)z^{-n} = 1 + z^{-1} + z^{-2} + z^{-3} + \dots + z^{-(N-1)} = \frac{1 - z^{-N}}{1 - z^{-1}}$$

$$W(\theta) = W(z)|_{z=e^{j\theta}} = \frac{1 - e^{-jN\theta}}{1 - e^{-j\theta}};$$

$$W(\theta) = \underbrace{e^{-j\frac{N-1}{2}\theta}}_{\text{Phase term}} \underbrace{\frac{\sin \frac{N\theta}{2}}{\sin \frac{\theta}{2}}}_{\text{Magnitude}}$$

If N is too small, E_m will fluctuate very rapidly depending on exact details of the waveform.

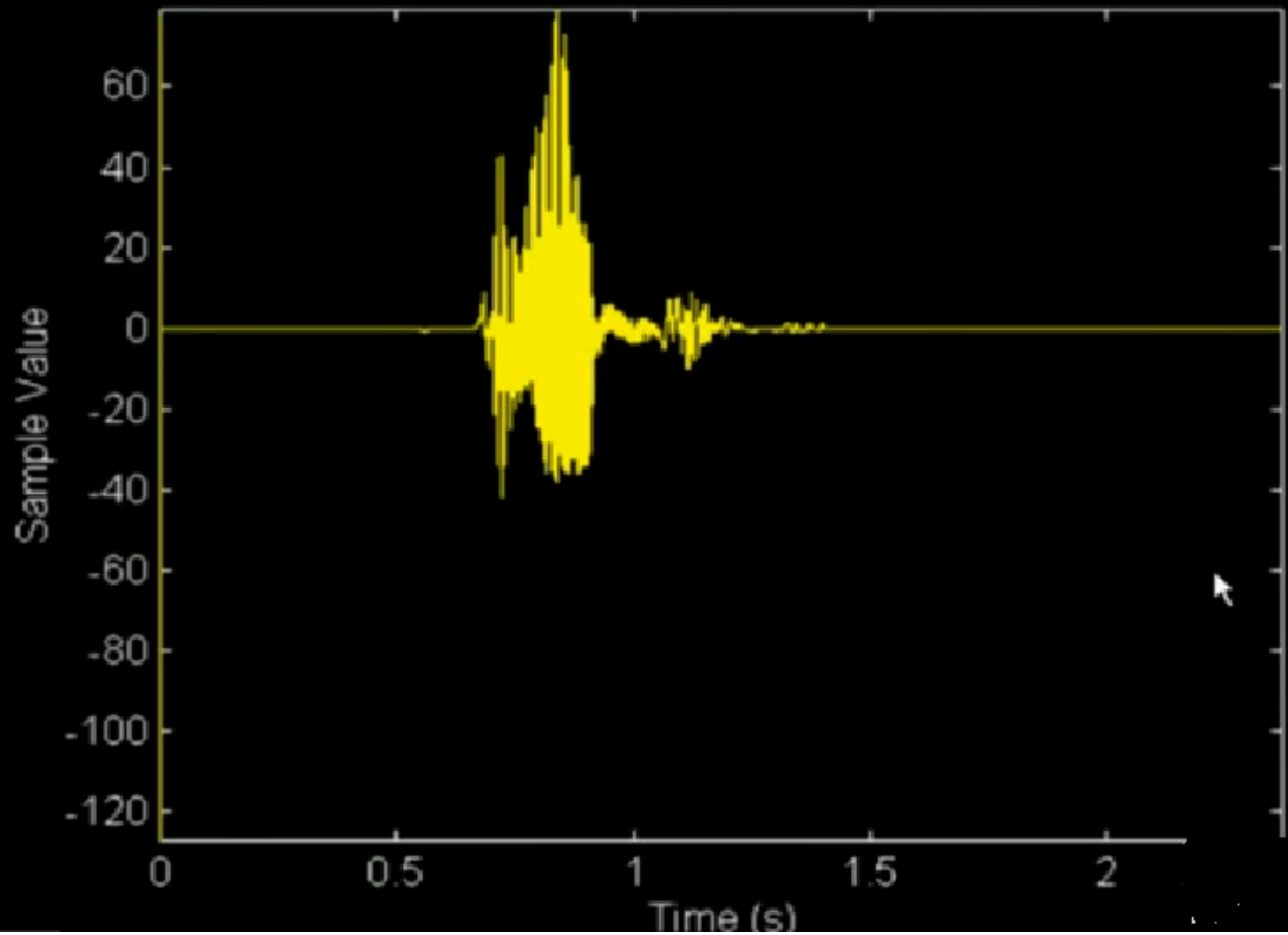


If N is too large, E_m will change very slowly and thus will not adequately reflect the changing properties of the speech signal.

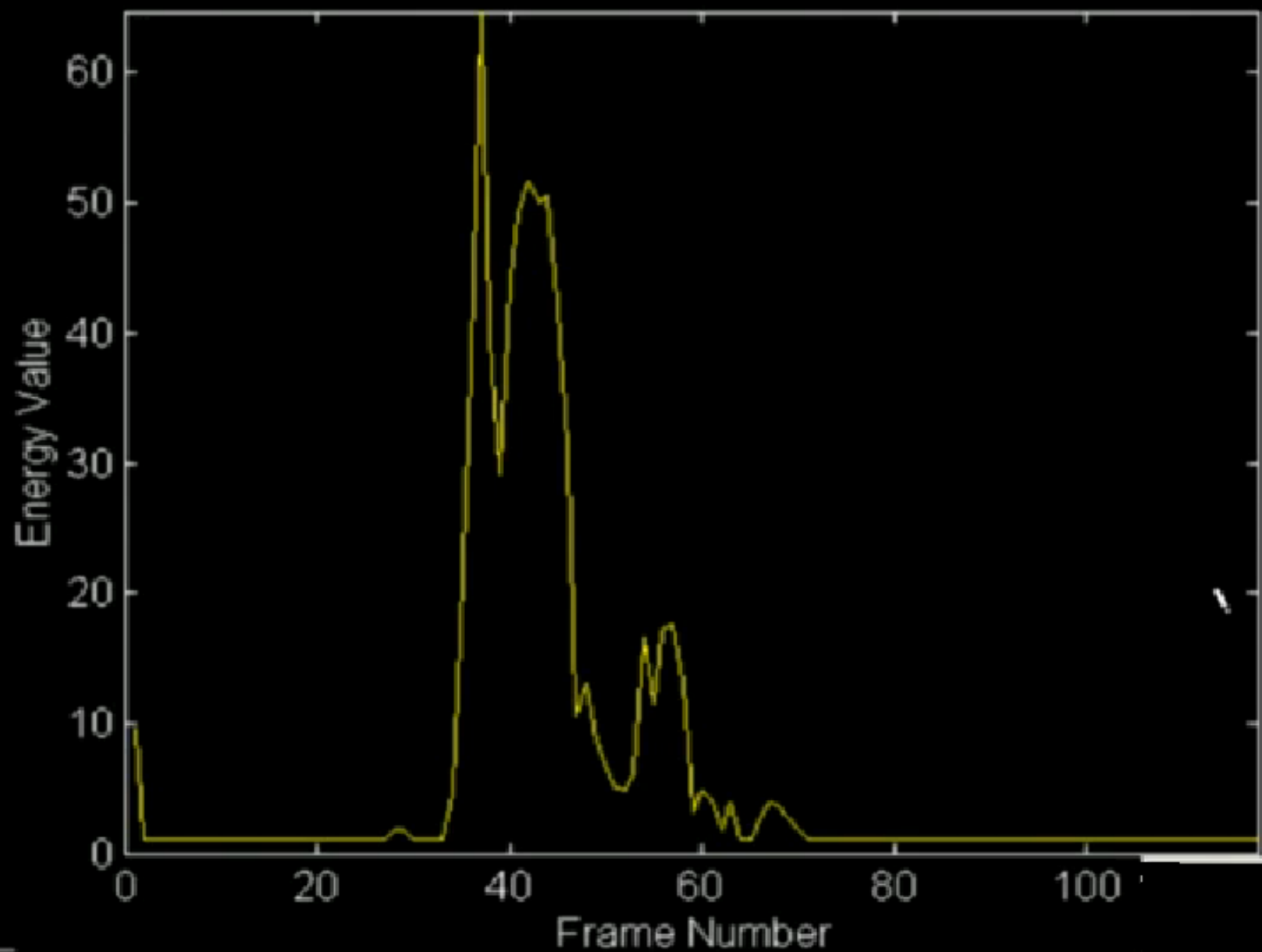
Choice of Window Size

- Unfortunately this implies that no single value of N is entirely satisfactory.
- A suitable practical choice for N is on the order of 100-200 samples for a 10 kHz sampling rate (10-20 ms duration)

Speech with DC offset removed



Short Time Average Energy Plot for Utterance



Note that a recursive lowpass filter $H(z)$ can also be used to calculate the short-time energy:

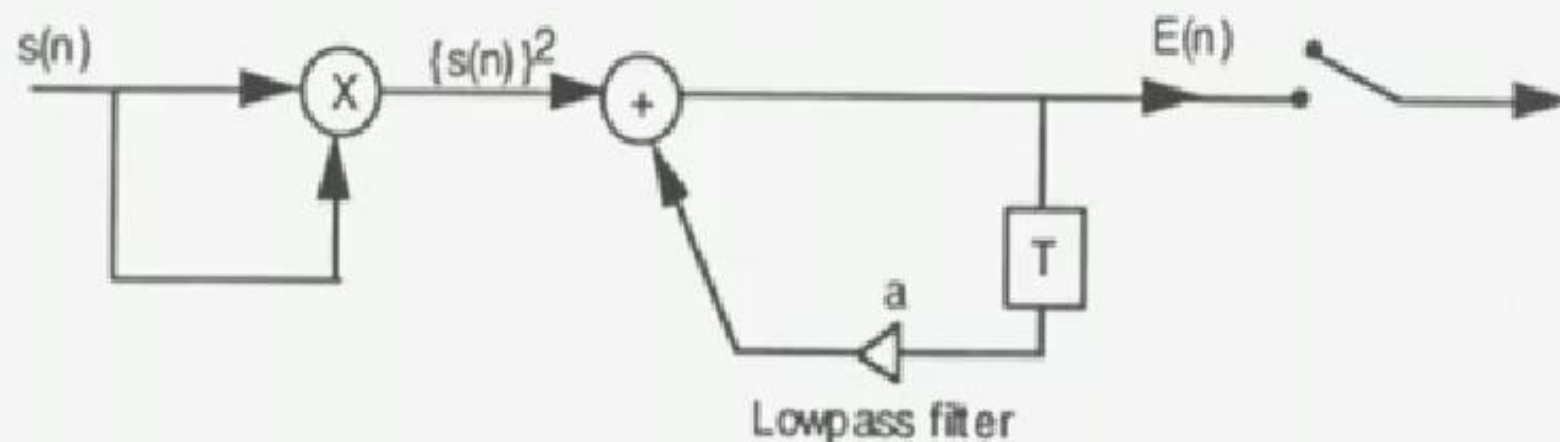
$$H(z) = \frac{1}{1 - az^{-1}} \quad 0 < a < 1$$

It can be easily verified that the frequency response $H(\theta)$ has the desired lowpass property. Such a filter can be implemented by a simple difference equation:

$$E(n) = a E(n-1) + [s(n)]^2$$

$E(n)$ is the energy at the time instant n

The structure for calculating the short-time energy recursively



The quantity $E(n)$ must be computed at each sample of input speech signal, even though a much lower sampling rate suffice.

The value 'a' can be calculated using

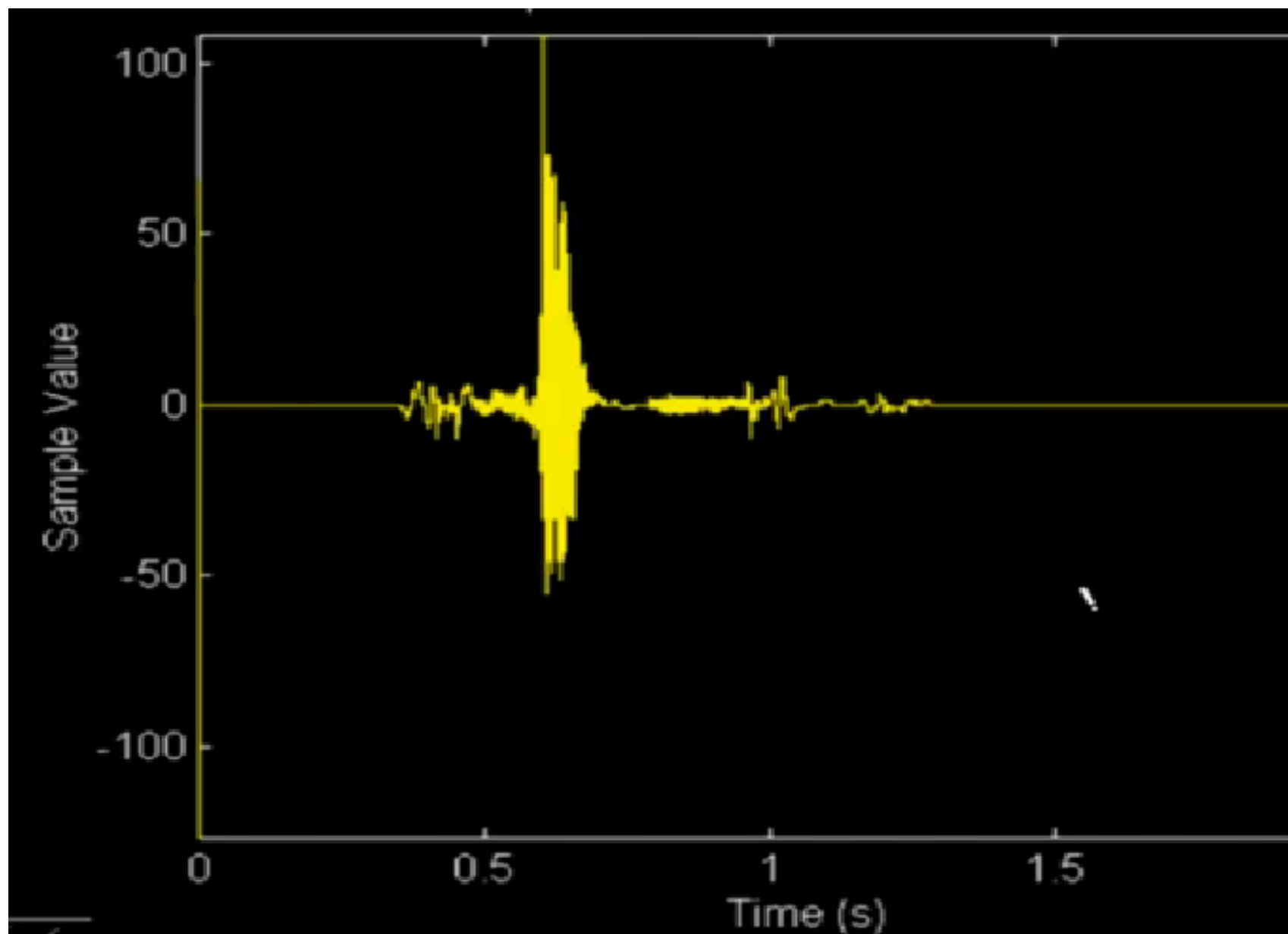
$$a = e^{(-f_c 2\pi / f_s)}$$

Short Time Zero Crossing Count

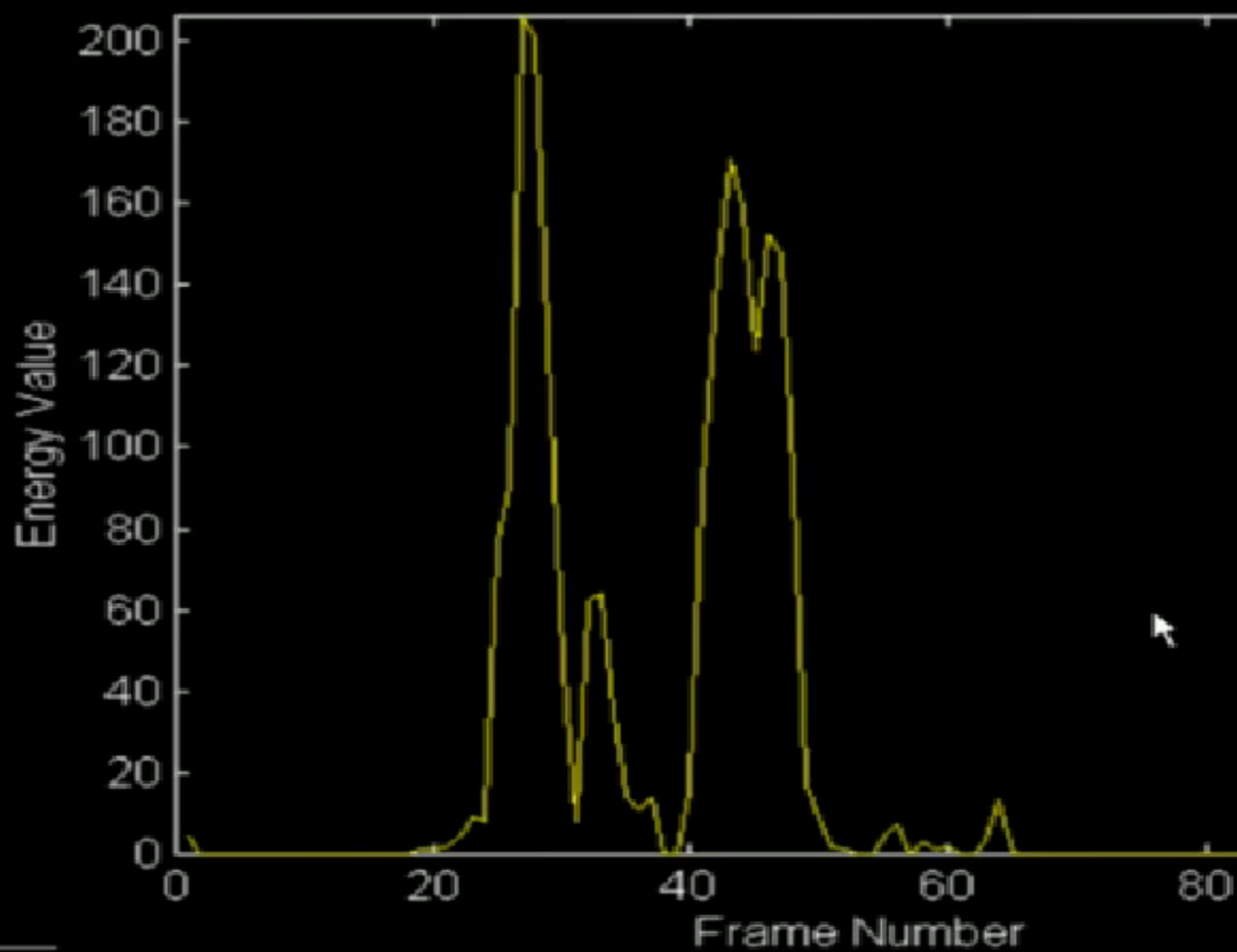
- The Short Time ZCC is calculated for a block of N samples of speech as

$$ZCC_i = \sum_{k=1}^{N-1} 0.5 | \text{sign}(s[k]) - \text{sign}(s[k-1]) |$$

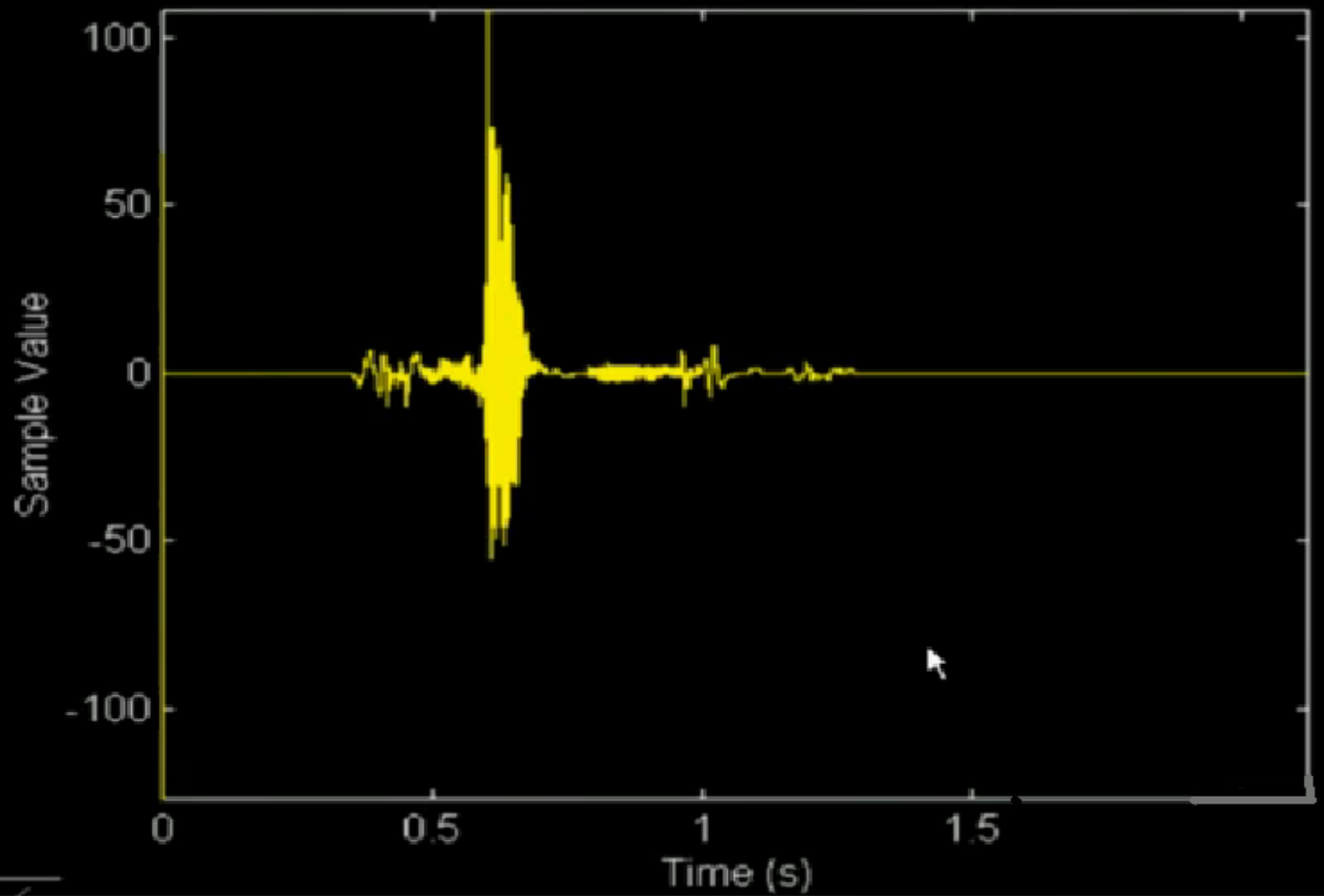
- The ZCC essentially counts how many times the signal crosses the time axis during the frame
 - It “reflects” the frequency content of the frame of speech
 - High ZCC implies high frequency
- It is essential that any constant DC offset is removed from the signal prior to ZCC calculation



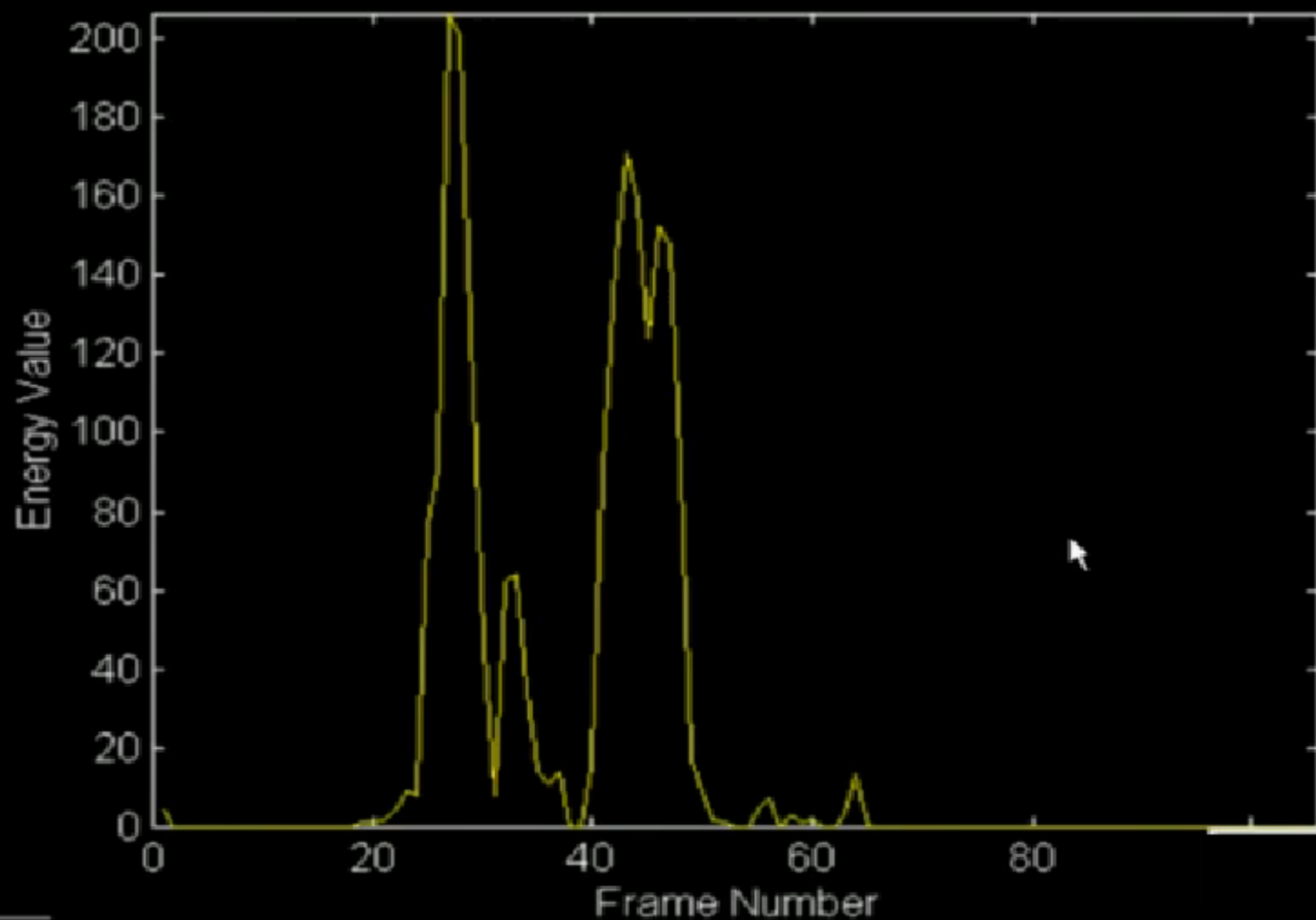
Short Time ZCC Plot for Utterance



Speech with DC offset removed



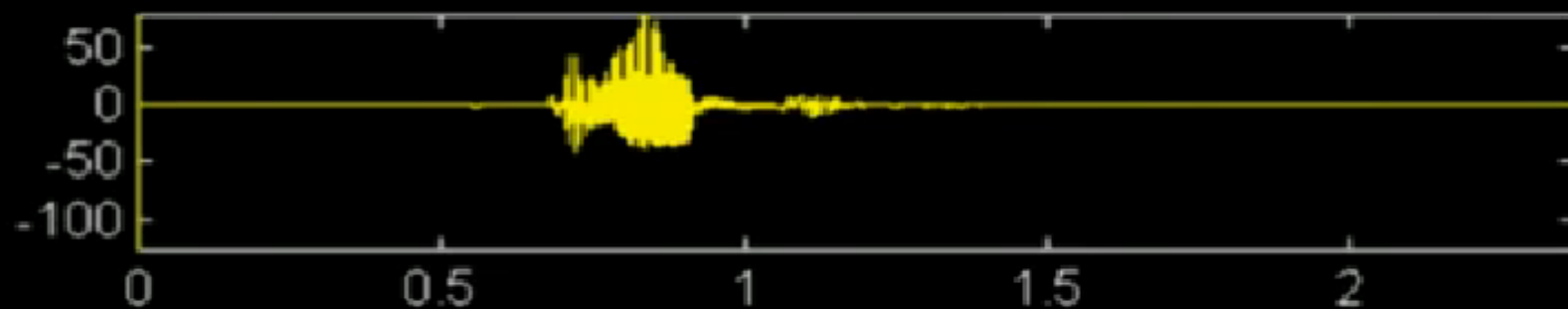
Short Time ZCC Plot for Utterance



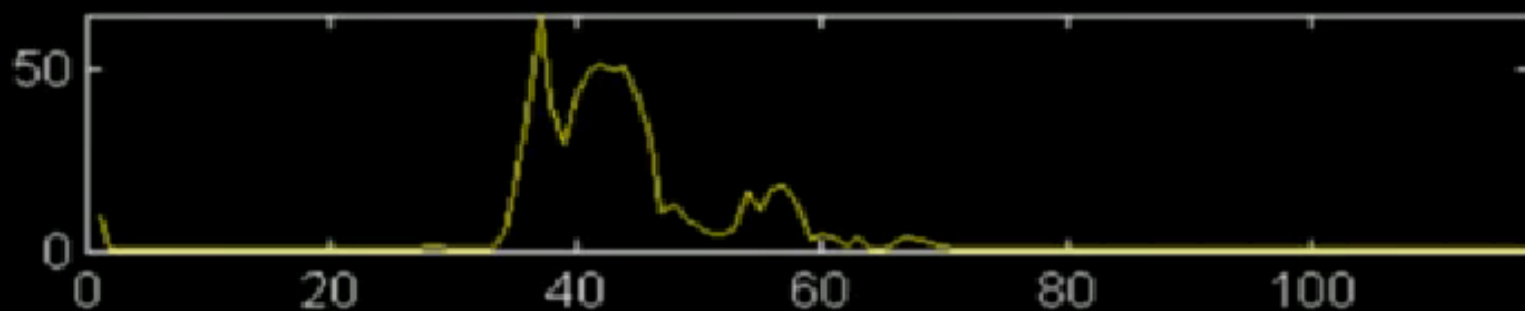
Uses of Energy and ZCC

- Short Time Energy and ZCC can form the basis for :
 - Automated speech “end point” detection
 - Needs to be able to operate with background noise
 - Needs to be able to ignore “short” background noises and intra-word silences (temporal aspects)
 - Voiced\Unvoiced speech detection
 - High Energy + Low ZCC – Voiced Speech
 - Low Energy + High ZCC – Unvoiced Speech
 - Parameters on which simple speech recognition\speaker verification\identification systems could be based

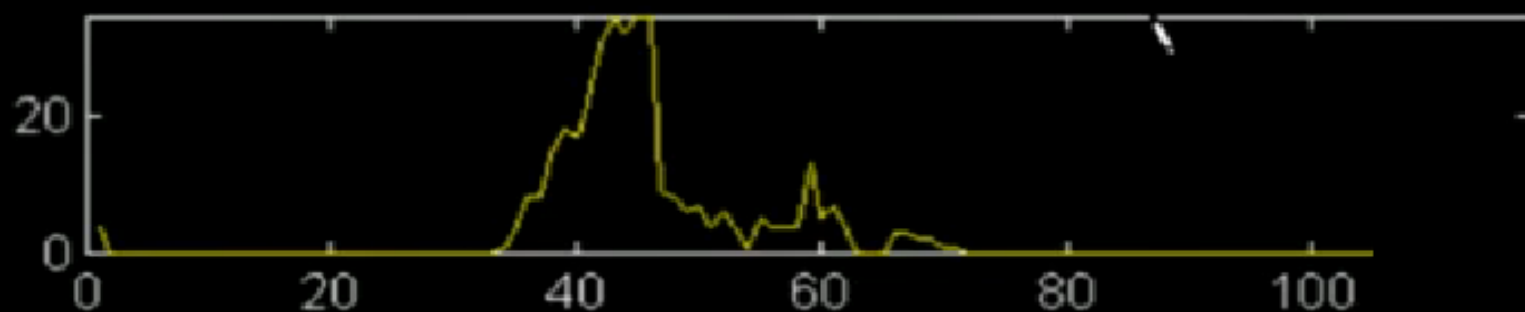
Speech Signal "ONE"



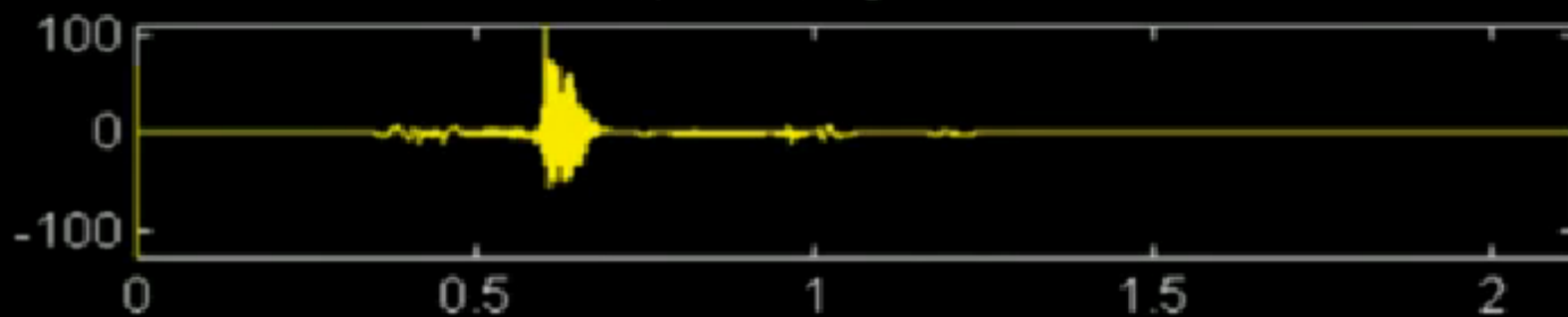
Short Time Energy



Short Time ZCC



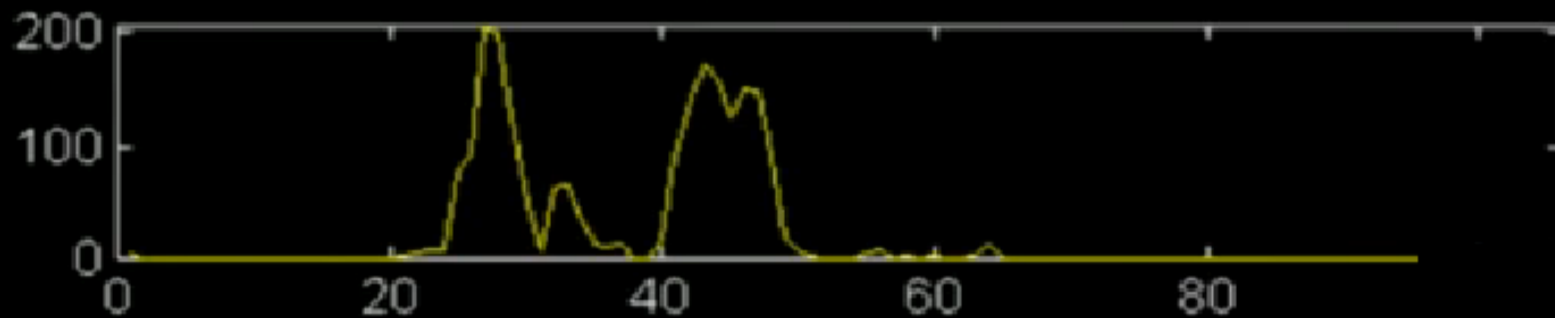
Speech Signal "SIX"



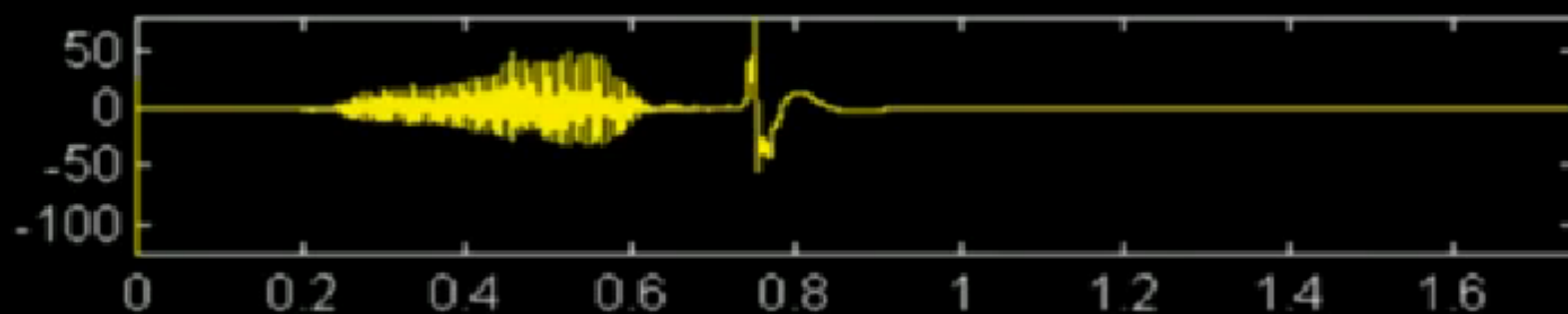
Short Time Energy



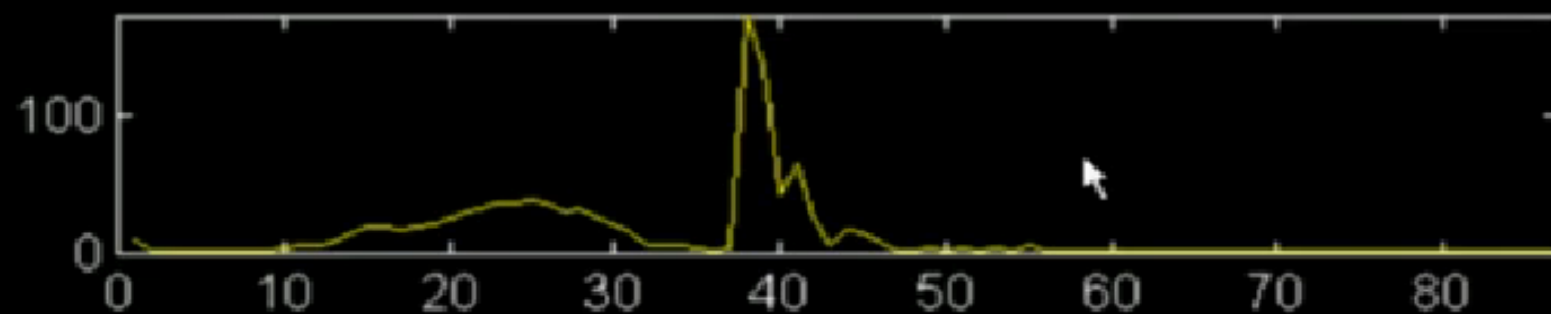
Short Time ZCC



Speech Signal "NINE"



Short Time Energy



Short Time ZCC

